27 October 2014

U.S. EPA Region 5 Report for the Statistical Analysis of Cadmium, Copper, Lead, Tin and Zinc Found Soil at and near the H. Kramer facility, Chicago, IL

USEPA FIELDS Group
John Canar, Environmental Scientist
Linda Jacobson, Research Associate
Chuck Roth, Life Scientist

Introduction

Soil samples were collected by the USEPA and its contractor, Weston Solutions, near the H. Kramer property as well as at locations up to a mile and a half away from the property. These samples were analyzed for metals by an accredited laboratory. The metals focused on for this study were Cadmium, Copper, Lead, Tin, and Zinc because these metals are more indicative of the metals present in H. Kramer airborne emissions. The purpose of these analyses was to investigate the similarities and differences in concentrations of Cadmium, Copper, Lead, Tin, and Zinc in soils on and near the H. Kramer property, the nearby Pilsen residential neighborhood, two local areas (Little Italy and Harrison Park (West)), and the USGS – Chicago Department of Environment surface metals sampling data (Kay et al., 2003).

Methods

Data sets

The USEPA-Weston samples consisted of grab and composites containing soil from up to five discrete locations on a given property. Soil samples were collected from the following depths: 0-6, 0-12, 6-12, 6-18, 6-24, and 18-24 inches below ground surface (bgs). The samples from the 0-6 inches bgs interval were used in these analyses. Samples were taken in front and back yards, alleys, and in soil areas with railroad tracks. The samples taken in gardens and drip zones were not used in this analysis due to garden soils potentially being amended, mixed and often imported, and drip zones being likely to contain Lead from Lead-based paint. Additionally, replicate samples and duplicate samples were also not used in this analysis. The samples were separated into seven areas called Railroad, Alley, Res1, Res2, Res3, Little Italy, and West (see Figure 1). Little Italy is considered the local reference area. Little Italy was selected as it was mostly crosswind/upwind from the H. Kramer smelter and, compared to the Pilsen-Kramer area, had a more limited industrial past and was similar in terms of age. Figure 2 is a representation of the historic wind rose for the Pilsen-Kramer area and environs. Note that "arms" in the figure represent the direction from which the wind blows; the lengths represent the proportion of the time the wind came from each direction (i.e., the frequency). Hence, for this wind rose, the predominant winds are from the west and the south. The Res1, Res2, and Res3 areas were created based on the spatial grouping of the USEPA's residential soil sampling locations and the prevalent wind directions (from 1928 to 2013). The three areas are presented in Figure 3. The wind directions were presented in Figure 2. The West area, i.e., near Harrison Park (see Figure 1) is also a potential local reference area although it may have been impacted by historic heavy-metal emitters that were located in that area.

Additional data used in these analyses included H. Kramer "on-site" and the "USGS" surface metal concentrations. The former data were taken from the "CRA Updated Focused Site Investigation Report Sept. 2007" created by Conestoga-Rovers and Associates. The latter were obtained from the report by Kay et al. (2003), i.e., the joint USGS – Chicago Department of Environment sampling event in 2000 and 2001.

Basic Statistics

The basic (descriptive) statistics were generated for the three near residential areas: Res1, Res2, and Res3. These areas were presented in Figure 3 above.

Multiple Comparisons

The comparison of metal levels (Cadmium, Copper, Lead, and Zinc) for each area and/or dataset was performed using a statistical procedure called ANOVA (analysis of variance). The metal levels' areas are shown in Figure 1; a total of eight areas. These levels were also compared to the USGS – Chicago Department of Environment (USGS) sampling results. (Tin was not used in these comparisons as the on-site data did not contain concentration values for Tin.)

The comparison of these areas was phrased in the form of a question: Is there a difference in metal levels in these areas? If metal levels in the Railroad, Alley, On-site, Res1, Res2, and/or Res3 were higher than those in Little Italy, West, or the Chicago area (the USGS – Chicago Department of Environment data) then this would indicate contamination. In order to answer this question, an ANOVA procedure is performed to test the hypothesis that the metal levels in each area are the same. Hence, one is testing whether Zinc levels, for instance, are the same for the Railroad, Alley, On-site, West, Res1, Res2, Res3, Little Italy, and the Chicago area. If that hypothesis is rejected, meaning that the levels of Zinc are not the same in these areas, then a multiple comparison procedure is performed. Since an ANOVA does not tell you which areas are different from each other, a multiple comparison procedure is performed to answer this question. The paragraph below explains how this is done using statistical software.

Since the data were not normally distributed for any of the metals (shown by the Shapiro-Wilk test; results not shown), and therefore violated the assumption of normality, the data were ranked to perform a nonparametric analysis. SAS® statistical software was used to compare the areas using one-way ANOVA on the ranked data with the general linear models (GLM) procedure. The Type III Sums of Squares result was used since the areas had an unbalanced number of samples. The Least Squares Means Tukey-Kramer Multiple Comparisons test was used to determine differences between the areas including the USGS dataset. The Least Squares Means Tukey-Kramer Multiple Comparisons test was selected because it accommodates unequal sample sizes and is the most robust test for pairwise comparisons (SAS, 2011).

Confidence Limits

As with the multiple comparison procedure, confidence limits were created for the ratio of Zinc to Lead in the eight areas and the USGS dataset. This ratio was found to be highest in the Railroad, Alley, and On-site samples compared to other areas at Pilsen.

Although the data were not normally distributed for the Zinc to Lead ratios (shown by the Shapiro-Wilk test; results not shown), and therefore violated the assumption of normality, the confidence limits were estimated parametrically and non-parametrically using the SAS® statistical software. In the majority of cases, there was little difference between the estimates.

Regression with distance

The SAS® statistical software was used to create simple linear regression models to predict Cadmium, Copper, Lead, Tin, and Zinc concentrations as a function of distance from the H. Kramer property's center. The statistical methods employed were drawn from SAS® literature and three regression texts: Statistical Methods in Water Resources, 1992; and Applied Regression Analysis and Other Multivariate Methods, 1978 and 1988.

The steps used to perform simple linear regression were:

1. Plot the data;
2. Compute the least squares regression statistics;
3. Examine adherence to the assumptions of regression using residual plots; and
4. Employ regression diagnostics (Helsel and Hirsch, 1992).

Results and Conclusions

Basic Statistics

The basic (descriptive) statistics for the three near residential areas are presented in Figure 4. The figures demonstrate a decline in metal levels as one gets farther away from the H. Kramer site going in the north, northeast, and east directions. Nonetheless there were still elevated Lead levels in Res 2 (median value of 930ppm) and in Res 3 (median value of 410ppm).

Multiple Comparisons

There was a significant difference between the eight areas and the USGS dataset for Cadmium, Copper, Lead, and Zinc (shown by one-way ANOVA on ranked data; results not shown). The Tukey-Kramer multiple comparison results for Lead are shown in Figure 5. A visual representation of the multiple comparisons for Cadmium, Copper, Lead, and Zinc is presented in Figure 6.

In Figure 6, the colored ovals represent areas with metal values that were not statistically different from each other. For example, the levels of Lead were not significantly different for samples from the RR, Alley, W, Res1, OS, and Res2. (Where RR is Railroad, Alley is Alley, W is West, Res1 is Res1, OS is on-site, and Res2 is Res2.) In contrast, the USGS and LI (Little Italy) area had significantly lower Lead levels than all of the above areas. But not significantly different Lead levels than in Res3. The blue oval, that overlaps both the grey and purple ovals, shows that although Res3 had statistically lower Lead levels than RR, Alley, W, and Res1, its levels were not significantly different than OS and Res2.

In general the figure demonstrates, when viewing from left to right, that Res1 and Res2 were not statistically different from each other. And, save for Copper, were not statistically different than the areas RR, Alley, OS (i.e., soils on and near the H. Kramer property). Lead samples from the West (i.e., Harrison Park area) were often not statistically different from Res1 and/or Res2 and some of the RR, Alley, OS areas. However, the elevated Lead levels in the West area are believed to be from a different source or sources, independent of H. Kramer. Additionally, the USGS dataset, Res3, and Little Italy often had significantly lower metal levels than all other areas. Additionally, these three areas were not statistically different from each other for all four metals.

Confidence Limits

The confidence limits, by area and the USGS dataset are shown in Figure 7. (A 95% confidence limit "means that if you took repeated random samples from a population and calculated the mean [or median] and confidence limits for each sample, the confidence interval for 95% of your samples would include the parametric mean [or median]" (McDonald, 2009). As can be seen in Figure 7, the confidence limits for the median Zinc to Lead ratios ("signatures") for the soils on and near the H. Kramer property (i.e., RR, Alley, On-site) overlap with that of Res1.

Statistically, the Zinc to Lead confidence limit for Res1 was no different than those for the RR, Alley, and On-site samples. This overlapping signature became less and less similar with Res2 and less so with Res3. The confidence limit for the "West" samples had a very different Zinc to Lead signature (the confidence limit) than the RR, Alley, On-site, Res1, and Res2 samples.

The importance of these confidence limits is to point out the overlap in median Zinc to Lead ratios observed in soil from areas adjacent to and on the H. Kramer property (i.e., Alley, RR, On-Site) and the impacted, near residential areas (Res1 and Res2). They also demonstrated the different contamination signature of the elevated metal levels in the West samples.

Regression with distance

The regression of metals levels (Cadmium, Copper, Lead, Tin, and Zinc) with distance from the H. Kramer site was statistically significant for each metal. Figures 8 through 12 show these relationships and associated statistical outputs. (The regression was performed on the natural log of the metals levels and distance in order to meet the assumptions of regression, specifically the homoscedasticity of residuals.) The slope for each of these regression equations was statistically significant and negative, indicating a decrease in metal concentrations with distance from H. Kramer. (See the "parameter estimate" for the LN_dist variable in the statistical output inset in each figure.) Additionally, these findings confirm the wind-borne conceptual site model. Moreover, these findings demonstrate that other potential sources, e.g., National Lead to the Northeast and Loewenthal to the East are not the source of these elevated metal levels. If the latter were true, metal levels would increase with distance from the site to these locations.

Summary

Overall, the EPA Fields Group's Statistical Analysis of Cadmium, Copper, Lead, Tin, and Zinc found at and near the H. Kramer facility indicates that H. Kramer is a significant contributor for elevated lead in residential surface soil in the RR/Alley, Res 1 and Res 2. However, the analyses could not conclude that there was lead contribution from H. Kramer in residential surface soils in Res 3. Further, Res 3 lead levels in surface soil indicate contributions from other industrial sources. Finally, the analysis indicated no apparent lead contribution in surface soil, from H. Kramer, in Harrison Park.

References

Helsel, D.R. and Hirsch R.M., Statistical Methods in Water Resources, Elsevier, Amsterdam, 1992.

Kay, R.T., Arnold, T.L., Cannon, W.F., Graham, D., Morton, E., and Biernet, R.

"Concentrations of Polynuclear Aromatic Hydrocarbons and Inorganic Constituents in Ambient Surface Soils, Chicago, Illinois: 2001-2002". United States Geological Survey and the Chicago Department of Environment. Water-Resources Investigations Report 03-4105. Urbana, Illinois. 2003.

Kleinbaum, D.G. and Kupper, L.L., Applied Regression Analysis and Other Multivariate Methods, Duxbury Press, Boston, Massachusetts, 1978.

Kleinbaum, D.G., Kupper, L.L., and Muller, K.E., Applied Regression Analysis and Other Multivariate Methods, Second Edition. PWS-Kent Publishing Company, Boston, Massachusetts, 1988.

McDonald, J.H., Handbook of Biological Statistics, Second edition. Sparky House Publishing, Baltimore, Maryland, 2009.

SAS Institute Inc., SAS/STAT® User's Guide, Version 9.2, Cary, NC: SAS Institute Inc., 2011. (The GLM Procedure, Multiple Comparisons)

Contact

Please contact the FIELDS Group via John Canar (canar.john@epa.gov) about this document.

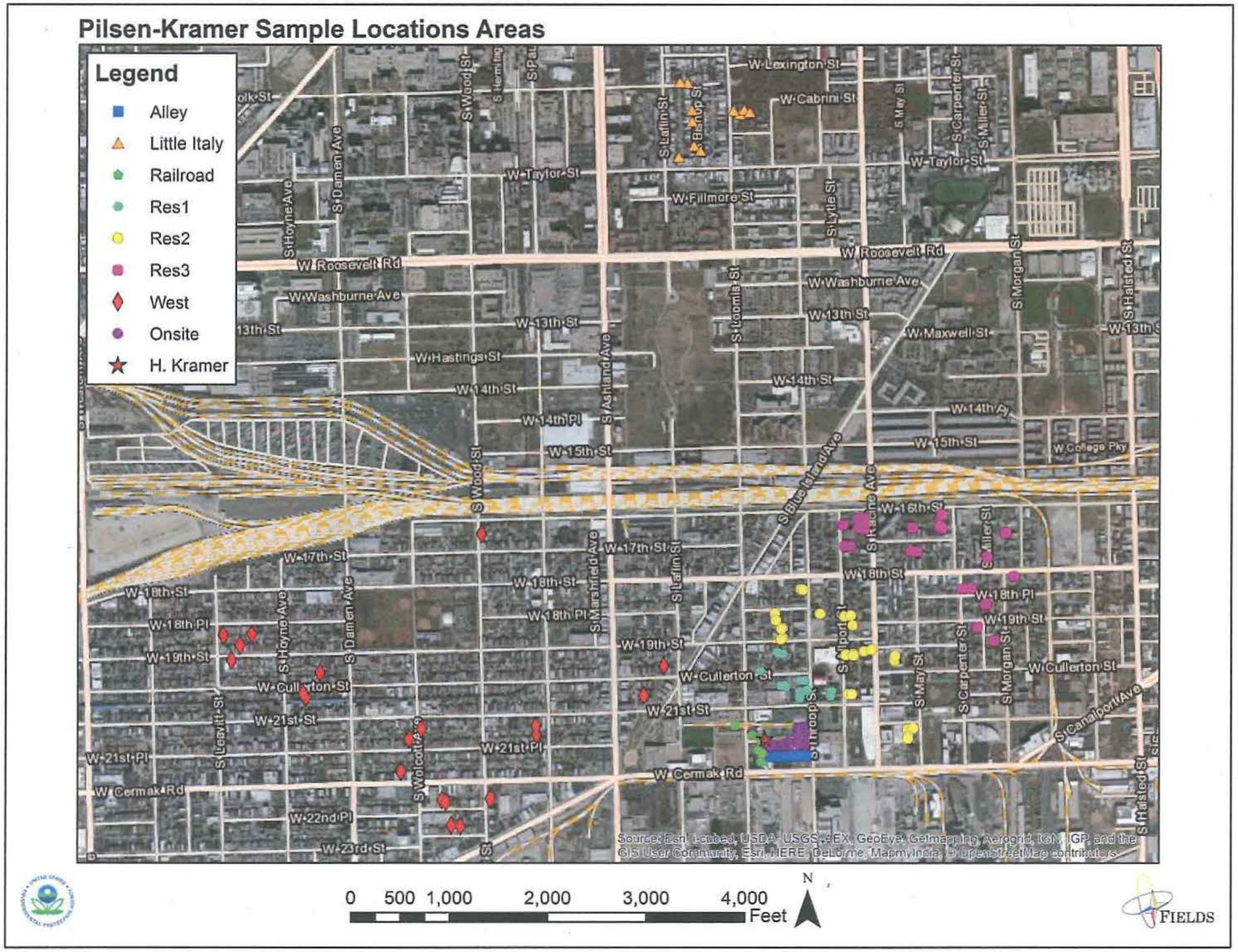


Figure 1: USEPA Sample locations and areas

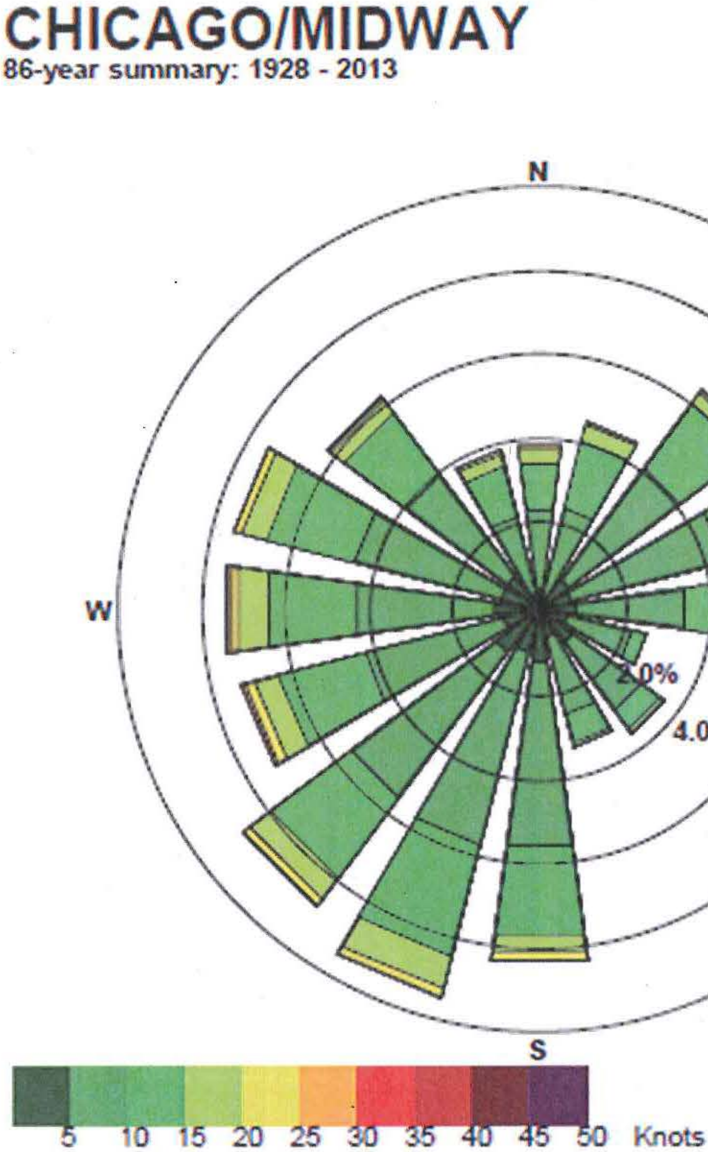


Figure 2: Windrose for the Pilsen-Kramer area and environs. Note that "arms" in the figure represent the direction from which the wind blows; the lengths represent the proportion of the time the wind came from each direction (i.e., the frequency). Hence, for this wind rose, the predominant winds are from the west and the south.

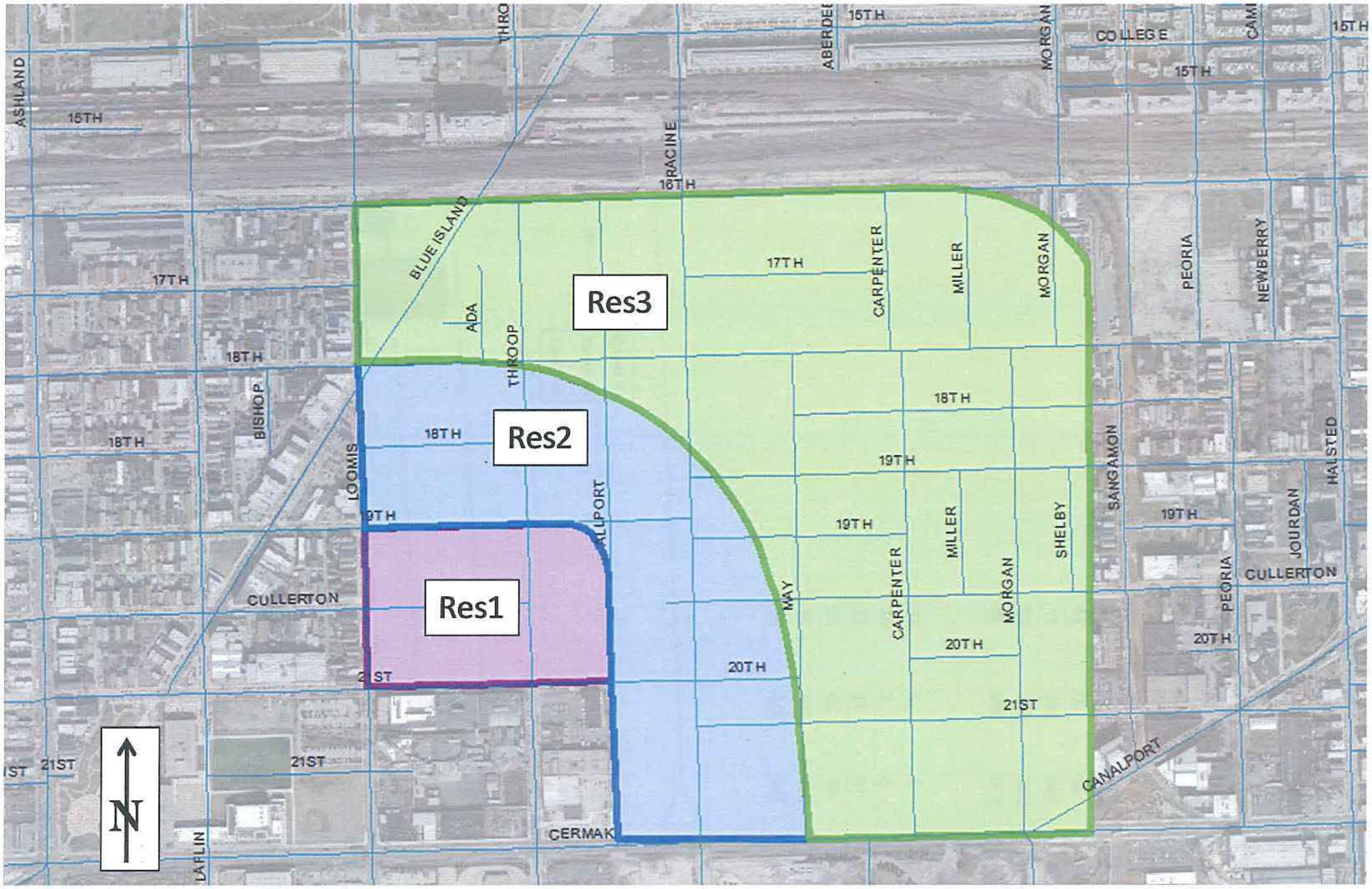


Figure 3: Near residential areas: Res1, Res2, and Res3.

Res1	mean	median	SD	N
Cd	7	7	5	14
Sb	11	3	14	14
Sn	80	59	60	14
Cu	535	425	447	14
Pb	1484	1850	904	14
Zn	2871	2650	2017	14
Res2				
Cd	4	4	3	27
Sb	4	5	1	27
Sn	38	28	30	27
Cu	207	190	133	27
Pb	1054	930	676	27
Zn	1320	970	906	29
Res3				
Cd	3	3	2	21
Sb	5	5	1	21
Sn	20	16	11	21
Cu	80	64	39	21
Pb	648	410	516	21
Zn	479	380	315	21

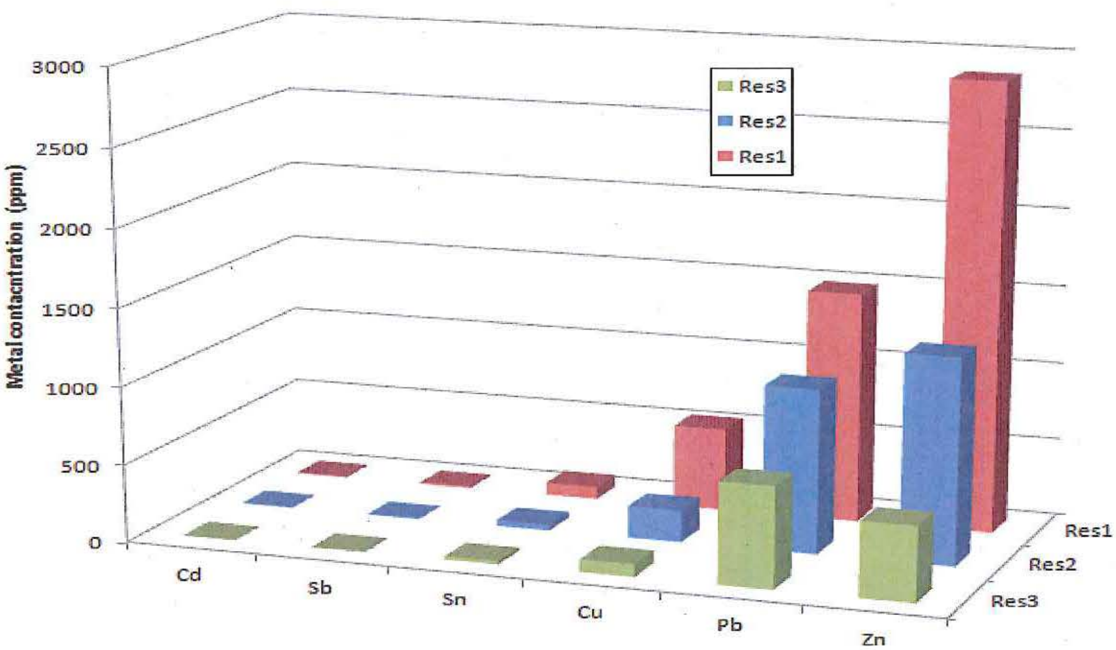


Figure 4: Basic Statistics

Proc GLM and post-hoc tests of differences; Ranks of Lead Levels by Area
Pilsen-Kramer Superfund Site
USEPA sampling (2012-2013), USGS-City of Chicago background data, and on-site data

The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

Area3	Lead_r LSMEAN	LSMEAN Number
Alley	134.318182	1
L_Italy	49.409091	2
On_site	109.000000	3
RR	156.500000	4
Res_1	127.785714	5
Res_2	107.629630	6
Res_3	82.000000	7
USGS	54.701754	8
West	128.857143	9

Least Squares Means for effect Area3
Pr > |t| for H0: LSMean(i)=LSMean(j)
Dependent Variable: Lead_r

i/j	1	2	3	4	5	6	7	8	9
1		0.0002	0.8458	0.9813	1.0000	0.6981	0.0274	<.0001	1.0000
2	0.0002		0.0133	<.0001	0.0002	0.0047	0.4877	1.0000	<.0001
3	0.8458	0.0133		0.3242	0.9549	1.0000	0.6142	0.0005	0.8974
4	0.9813	<.0001	0.3242		0.8964	0.2041	0.0055	<.0001	0.8883
5	1.0000	0.0002	0.9549	0.8964		0.8737	0.0477	<.0001	1.0000
6	0.6981	0.0047	1.0000	0.2041	0.8737		0.4792	<.0001	0.7233
7	0.0274	0.4877	0.6142	0.0055	0.0477	0.4792		0.2182	0.0116
8	<.0001	1.0000	0.0005	<.0001	<.0001	<.0001	0.2182		<.0001
9	1.0000	<.0001	0.8974	0.8883	1.0000	0.7233	0.0116	<.0001	

Figure 5: Tukey-Kramer multiple comparison results.

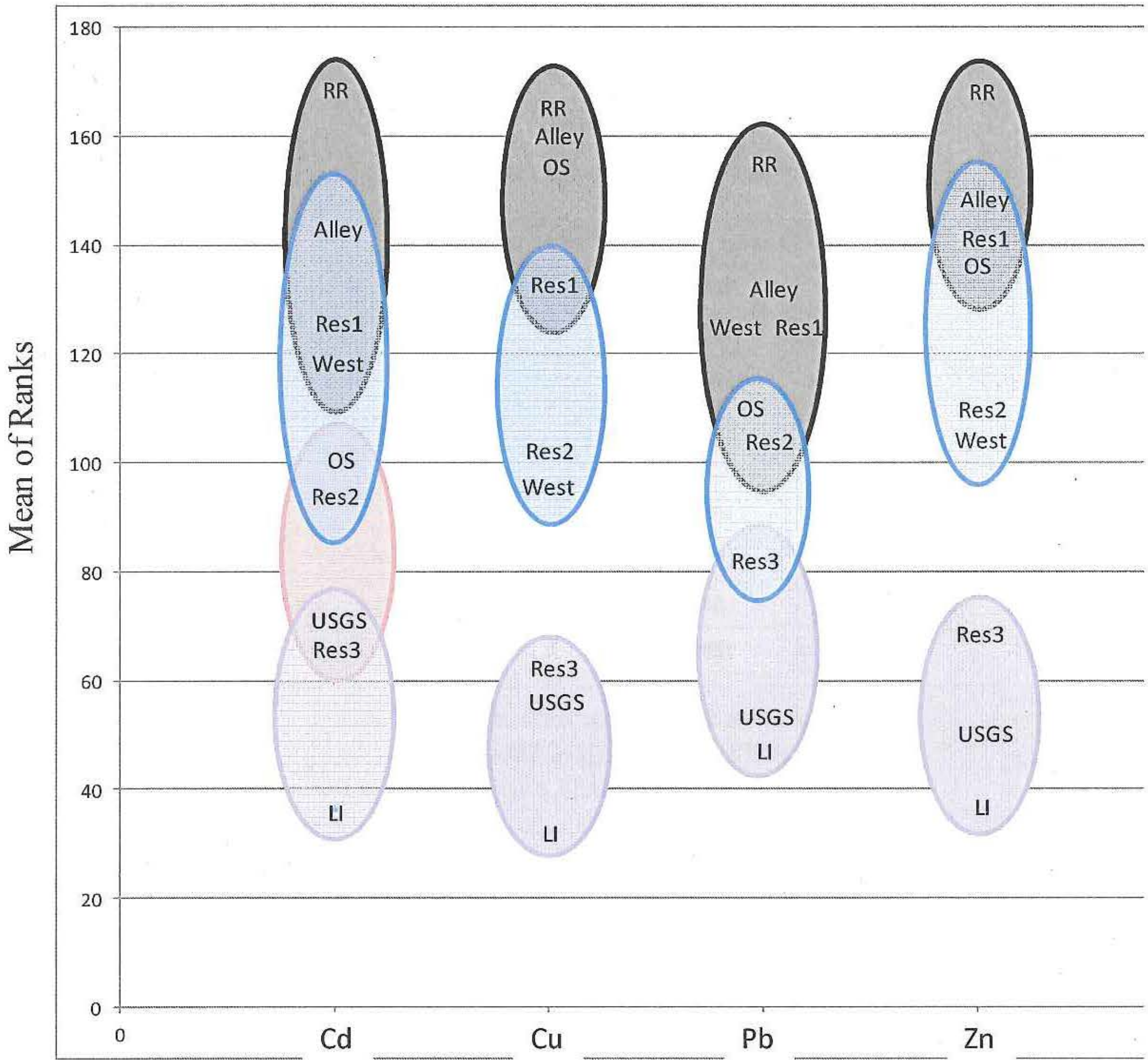


Figure 6: Cumulative schematic of the multiple comparisons by area and metal. Where LI is Little Italy, USGS is the USGS – Chicago Department of Environment dataset, Res1, Res2, and Res3 are as defined before, W is West, OS is on-site, Alley is Alley, and RR is Railroad. Where the Y-axis is the "lsmeans" value for each metal and dataset (the mean of the ranked values). Areas in the same colored ovals are not statistically different from each other; areas in different colored ovals are statistically different from each other.

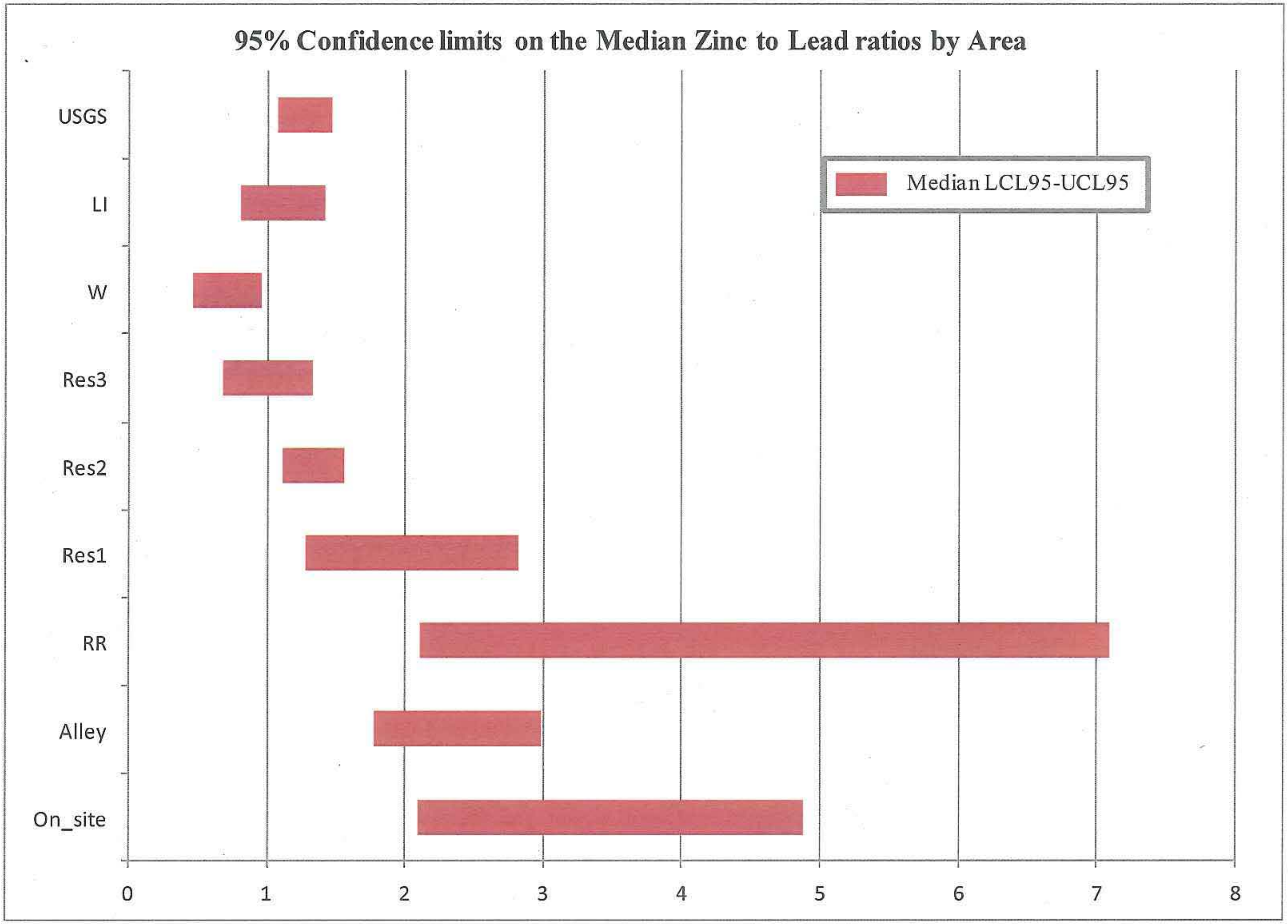
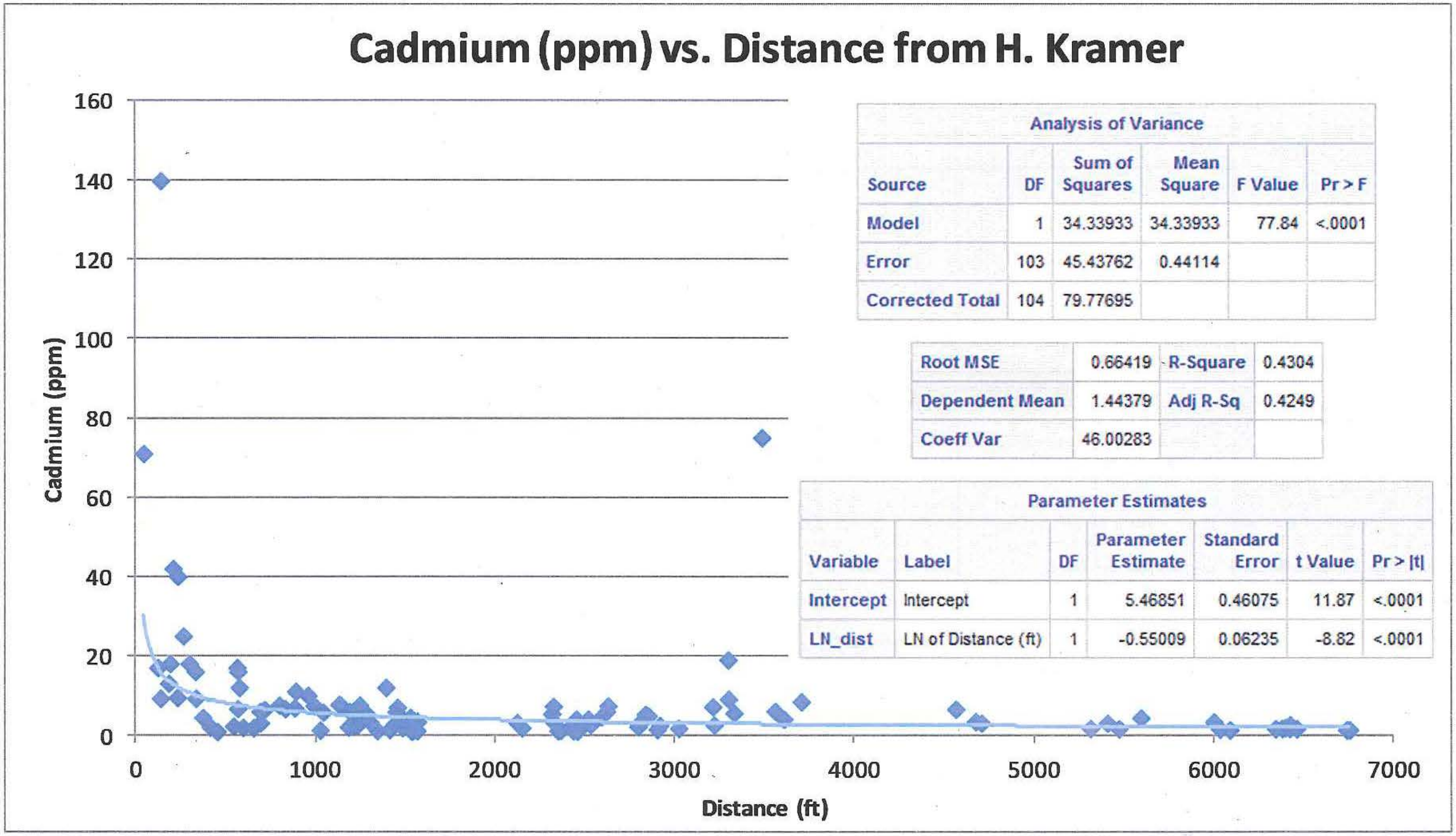


Figure 7: Confidence limits for the median Zinc to Lead ratio.

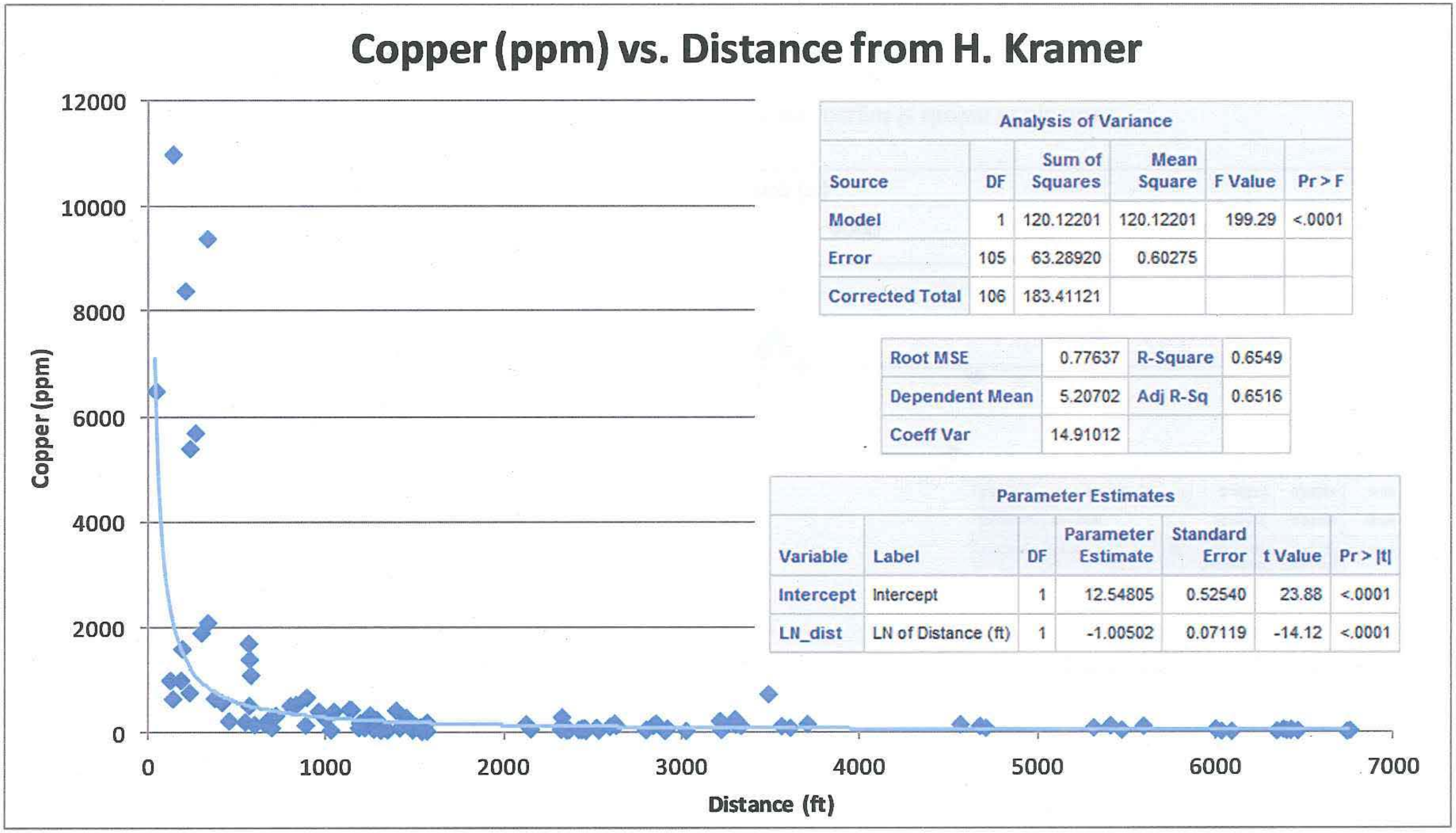


Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	34.33933	34.33933	77.84	<.0001
Error	103	45.43762	0.44114		
Corrected Total	104	79.77695			

Root MSE	0.66419	R-Square	0.4304
Dependent Mean	1.44379	Adj R-Sq	0.4249
Coeff Var	46.00283		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	5.46851	0.46075	11.87	<.0001
LN_dist	LN of Distance (ft)	1	-0.55009	0.06235	-8.82	<.0001

Figure 8: Cadmium levels as a function of distance from H. Kramer. Statistical output is shown in the inset.



Analysis of Variance

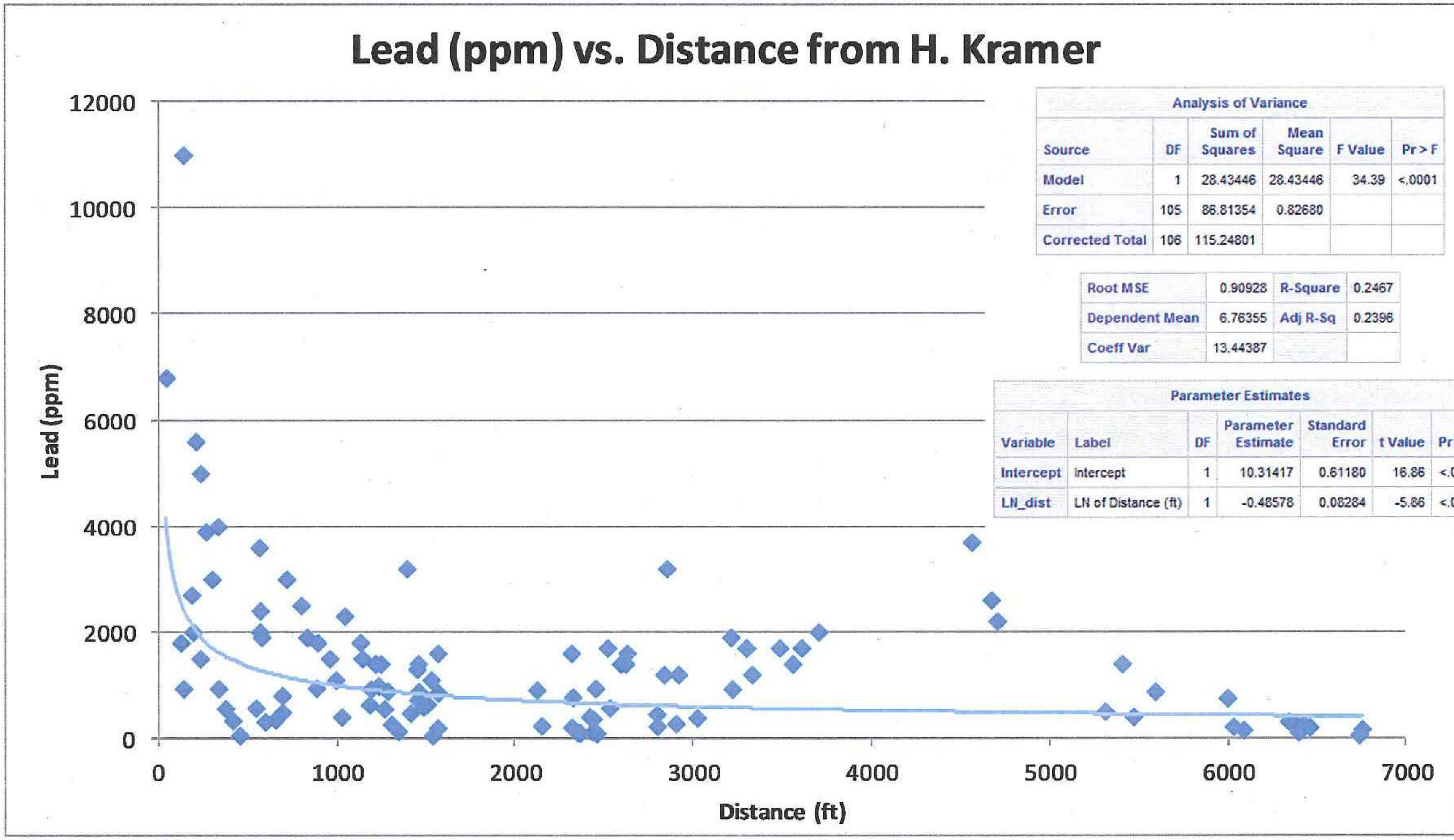
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	120.12201	120.12201	199.29	<.0001
Error	105	63.28920	0.60275		
Corrected Total	106	183.41121			

Root MSE	0.77637	R-Square	0.6549
Dependent Mean	5.20702	Adj R-Sq	0.6516
Coeff Var	14.91012		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	12.54805	0.52540	23.88	<.0001
LN_dist	LN of Distance (ft)	1	-1.00502	0.07119	-14.12	<.0001

Figure 9: Copper levels as a function of distance from H. Kramer. Statistical output is shown in the inset.

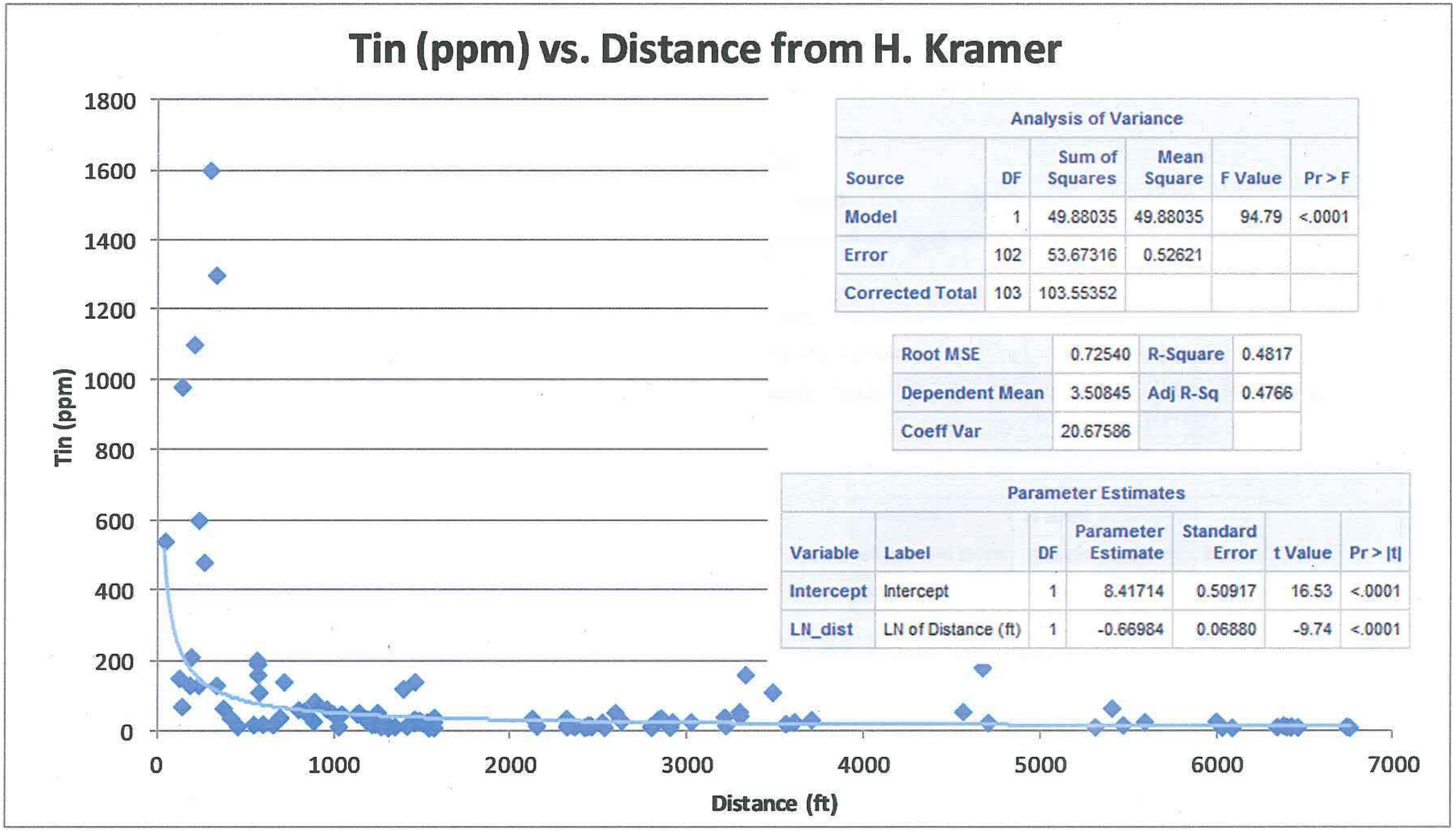


Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	28.43446	28.43446	34.39	<.0001
Error	105	86.81354	0.82680		
Corrected Total	106	115.24801			

Root MSE	0.90928	R-Square	0.2467
Dependent Mean	6.76355	Adj R-Sq	0.2396
Coeff Var	13.44387		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	10.31417	0.61180	16.86	<.0001
LN_dist	LN of Distance (ft)	1	-0.48578	0.08284	-5.86	<.0001

Figure 10: Lead levels as a function of distance from H. Kramer. Statistical output is shown in the inset.

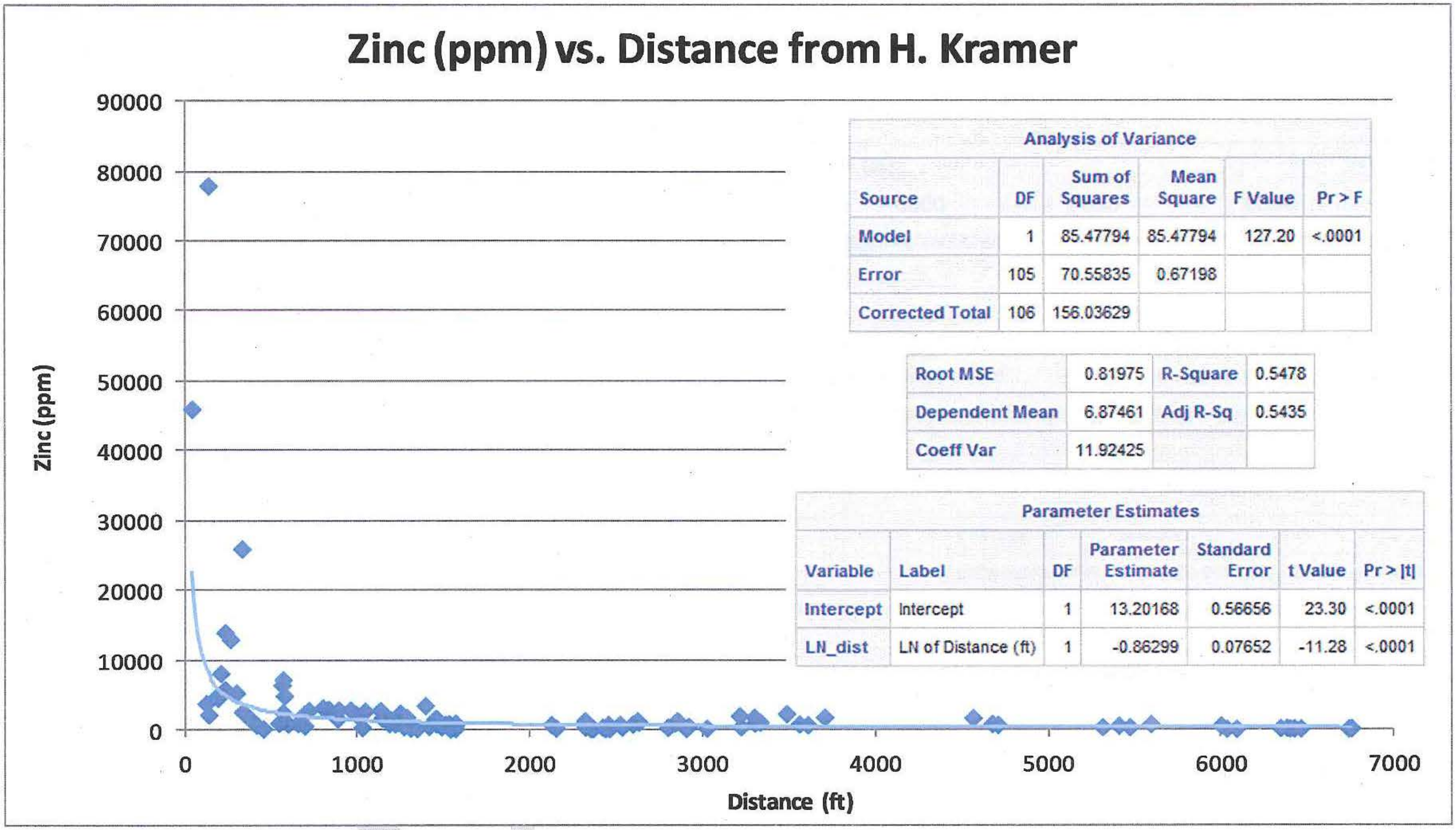


Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	49.88035	49.88035	94.79	<.0001
Error	102	53.67316	0.52621		
Corrected Total	103	103.55352			

Root MSE	0.72540	R-Square	0.4817
Dependent Mean	3.50845	Adj R-Sq	0.4766
Coeff Var	20.67586		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	8.41714	0.50917	16.53	<.0001
LN_dist	LN of Distance (ft)	1	-0.66984	0.06880	-9.74	<.0001

Figure 11: Tin levels as a function of distance from H. Kramer. Statistical output is shown in the inset.



Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	85.47794	85.47794	127.20	<.0001
Error	105	70.55835	0.67198		
Corrected Total	106	156.03629			

Root MSE	0.81975	R-Square	0.5478
Dependent Mean	6.87461	Adj R-Sq	0.5435
Coeff Var	11.92425		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	13.20168	0.56656	23.30	<.0001
LN_dist	LN of Distance (ft)	1	-0.86299	0.07652	-11.28	<.0001

Figure 12: Zinc levels as a function of distance from H. Kramer. Statistical output is shown in the inset.